Erschienen in: LDV-INFO 1996. Informationsschrift der Arbeitsstelle Linguistische Datenverarbeitung. - Mannheim: IDS, 1996. S. 112-122. (LDV-INFO 8)

Franck Bodmer

Aspekte der Abfragekomponente von COSMAS-II

0. Einleitung

Die wesentlichen Konzepte der Textdatenbank COSMAS-II in bezug auf die graphikorientierte Abfragekomponente sollen im folgenden beschrieben und einige davon etwas genauer erläutert werden. COSMAS-II ist ein Nachfolgesystem von COSMAS-I, das am IDS seit 1992 im Einsatz steht. Es befindet sich zur Zeit noch in der Entwicklungsphase und ermöglicht heben einer erhöhten Flexibilität und Benutzerfreundlichkeit der Abfragekomponente als wichtige Erneuerung den Einbezug von SGML-Daten (siehe Kap. 1). Dadurch erschließt das neue System das Abfragen einer Vielfalt von dokument- und domainspezifischen Textannotationen, mitunter linguistischen Annotationen.

Die Abbildung 1 stellt die wichtigsten Bestandteile des Abfragesystems von COSMAS-II dar, welche nun zur Sprache kommen sollen. Andere Komponenten wie Datenzugriff usw. werden hier nicht behandelt. Der graphische Suchanfrageneditor bzw. die graphische Form der Suchanfrage soll in diesem Komplex eine zentrale Stellung in bezug auf die Formulierung, Manipulation, Darstellung, Flexibilität, Mehrschichtigkeit und Umgebung der Suchanfragen einnehmen. Diese werden dann einem Datenbankzugriffsmodul, auf das nicht weiter eingegangen werden soll, überreicht. Die Ergebnismengen werden einerseits dem Editor für weitere Suchoperationen wieder zur Verfügung gestellt, andererseits einer Darstellungskomponente weitergereicht. Die Darstellung kann verschiedenartig sein, weil SGML-Daten miteingebracht werden können, weil zwischen verschiedenen Informationsebenen gewählt werden kann oder weil statistische Werkzeuge zur Modifikation oder Reorganisation der Ergebnisse verwendet werden können. Die graphische Form der Suchanfrage eignet sich ebenfalls als anschaulicher Darstellungsträger von z.B. Zwischenergebnisstatistiken. Schließlich besteht auch eine interessante Interaktion zwischen den Ergebnismengen und dem Korpusdefinitionsteil von COSMAS-II.

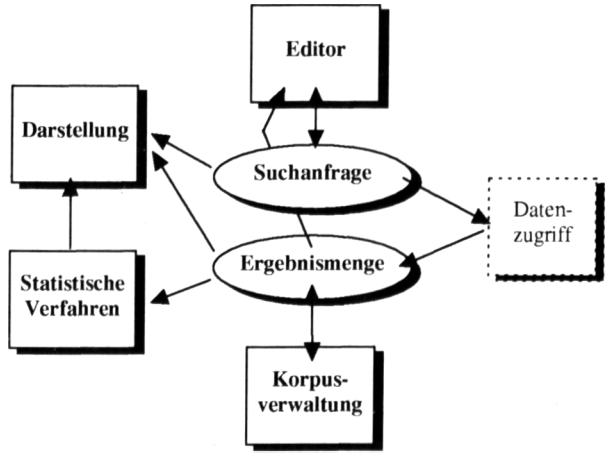

Abbildung 1: Bestandteile der Suchanfragekomponente

Die Abbildungen, die im folgenden abgedruckt sind, sollen die erläuterten Konzepte veranschaulichen, sie entstammen aber nicht der im Entstehen begriffenen Benutzeroberfläche (MS-Windows und X-Windows). Ähnliches gilt auch für die Beispiele zu der Suchanfragesyntax, die möglichst selbsterklärend sein sollen und auf Details der definitiven Syntax verzichten.

1. Die COSMAS-II-Textdatenbank

Bevor wir auf die Konzepte der Abfragekomponenten eingehen, sollen in diesem Kapitel kurz die Funktionen von COSMAS-II vorgestellt werden (mehr über das COSMAS-I-System kann in al-Wadi, 1994 nachgeschlagen werden).

COSMAS-II ist eine Text- oder Korpusdatenbank, die speziell für große Datenmengen (einige Hundert Millionen Wörter) konzipiert ist. Sie ist von einer Bausteinphilosophie durchzogen, die es erlaubt, um einen zentralen Datenverwaltungskern herum verschiedene linguistische Verfahren in Form von Programmen einzufügen, die z.B. die Texte anreichern können oder die Auswertung von großen Datenauszügen unterstützen. Beispiele von Anreicherungswerkzeugen sind ein integrierter Worteinheiten- und Satzsegmentierer und ein deutscher Lemmatisierer (für Wortstamm, Affixe und Wortbildungsmorpheme). Als Beispiel eines Auswertungswerkzeugs seien integrierte statistische Analysen genannt..

Der Datenzugriff erfolgt über eine Abfragesprache, die Textstellen anhand von Sprachmustern und deren Kombinationen extrahiert und bibliographisch dokumentiert. Dabei können die angereicherten Daten in den Sprachmustern mitspezifiziert werden, um genauere Muster zu formulieren. Das Teilkorpuskonzept erlaubt, durch eine gezielte Auswahl von Dokumenten gezielte Untersuchungen ausführen zu können (*Monitoring* von Zeitausschnitten, Sprachtypen, Dokumenttypen usw.).

Mit der Erweiterung der Textdatenbank durch SGML-Strukturen erhält der COSMAS-II-Benutzer die Möglichkeit, vielseitig anwendbare Textannotationen abzuspeichern und abzufragen. SGML (Standard General Markup Language, siehe z.B. Bryan, 1988) ist ein standardisierter Formalismus zur Beschreibung von Dokumentstrukturen (wie *Kapitel*, *Titel*, *Zitate* usw.), Dokumentlayout (wie typographische Informationen, Querverweise, Aufzählungen usw.), Dokumentinformationen (wie bibliographische Informationen usw.) und Textannotationen (wie syntaktische Kategorien, Wortlemmata, Wortsegmentbeschriftung usw.) und erleichert außerdem den Transfer von plattformabhängigen textuellen Parametern (wie Alphabet, Zeichensätze usw.).

SGML-Strukturen, die in COSMAS-II eingespeist werden, sollen somit zum Teil auf eine einfache Art direkt zugänglich gemacht werden, da sie zum Beispiel zusätzlich erlauben, Suchanfragen in einem bestimmten Kontext zu formulieren. Auf der anderen Seite erhöht sich dadurch die Komplexität der Dokumente, und zusätzliche Kenntnisse wie die Beschreibung der SGML-Dokumentstrukturdefinitionen (der sogenannten DTD: Document Type Definition) müßten unter Umständen erworben werden. Da, wo aber SGML als Mittel benutzt wird, um z.B. den standardisierten Transfer von extern angereichertem Material zu kodieren, soll seine Komplexität für den Benutzer möglichst transparent bleiben. Dies soll in COSMAS-II mit dem graphischen Suchanfrageneditor realisiert werden.

Neu in COSMAS-II soll auch die Teilkorpusbildung über bibliographische Suchanfragen möglich sein, da diese Art von Daten im SGML-Kopfteil der Dokumente mitgeliefert und somit in gleicher Weise wie der restliche Textteil abgefragt werden kann.

SGML ist natürlich nicht der einzige Formalismus, der zu diesem Zweck benutzt werden könnte, doch seine Verbreitung als Standard (ISO 8879) und sein Einbezug in der Ausarbeitung von TEI (Text Encoding and Interchange: eine Sammlung von Richtlinien und Empfehlungen für die Kodierung von textuellem Material aller Art wie geschriebene und gesprochene Korpora, linguistische Annotationen, Textparallelisierung usw., nachzuschlagen in Burnard & Sperberg-McQueen, 1994) machen ihn für COSMAS-II besonders angebracht und wertvoll.

2. Konzepte der Abfragekomponente von COSMAS-II

Wir möchten nun in diesem Kapitel einige der Konzepte darlegen, die uns bei der Konzeption der neuen Abfragekomponente geleitet haben. Die meisten dieser Konzepte werden in den nachfolgenden Kapiteln wieder aufgenommen und veranschaulicht.

Der Ausgangspunkt des Abfragesystems ist natürlich die Abfragesprache. Um der erhöhten Komplexität der abzufragenden Daten Rechnung zu tragen, soll in COSMAS-II die Möglichkeit bestehen, Suchanfragen auf verschiedenen Komplexitätsebenen zu formulieren. Einerseits existiert eine einfache Befehlszeile, in die ähnlich wie in COSMAS-I eine beliebige Suchanfrage eingegeben und abgeschickt werden kann. Dieser Modus ist erfahrenen Benutzern vorbehalten, welche Abfragesyntax und SGML-Strukturen usw. genau kennen und vielleicht spezielle Suchanfragen formulieren möchten, die im graphischen Modus nicht möglich sind.

Ein zweiter Modus ist durch einen flexiblen *graphischen Anfrageeditor* gegeben. Dieser hat zum Ziel, Suchanfragen graphisch aufzubauen, um ausgehend von ihnen alle weiteren Arbeiten vorzunehmen. Eine Leiste stellt eine Reihe von Suchoperatoren als *symbolische Kästchen* zur Verfügung, welche in einer beliebigen Reihenfolge zusammengestellt werden können. Z.B. kann auf diese Art eine Suchanfrage zuerst *top-down* als *Gerüst* aufgebaut werden, bevor in einem zweiten Schritt die noch leeren Operanden mit konkreten Suchspezifikationen ausgefüllt werden. Die vervollständigte Suchanfrage wird danach ausgeführt und steht als *Ergebnissymbol* zur Verfügung, welches entweder durch Anklicken zur Ergebnispräsentation geführt wird oder für die Formulierung von weiteren Suchanfragen wieder eingesetzt werden kann.

Baut ein Benutzer eine komplizierte Suchanfrage auf, die er unter Änderung einiger weniger Suchoperanden immer wieder verwenden möchte, so kann er veranlassen, daß die graphische Darstellung als neues Kästchen in die Liste der schon vorhandenen eingereiht wird. Die beim ersten Start von COSMAS-II vorhandenen Kästchen wollen wir die *Systemkästchen* nennen, die vom Benutzer hinzugefügten die *Benutzerkästchen*. Die Systemkästchen sind also diejenigen, aus denen sich alle anderen Suchanfragen aufbauen lassen. Ein zweiter Typ von Systemkästchen läßt sich in diesem System auch so konstruieren, daß ein Sachkundiger eine Suchanfrage aufbaut und ihr eine *Maske* überstülpt, welche nur so viele Operanden besitzt, wie in der Suchanfrage offene (freiwählbare) Operanden vorhanden sind. Diese Maske läßt sich dann verriegeln und ebenfalls zu den schon bestehenden Kästchen einreihen. Bei der Verwendung verhält sie sich genau wie ein

anderes Kästchen (mit einem Operator, x-beliebigen Operanden und einem dokumentierten Verhalten, dessen versteckte Komplexität den Benutzer nicht zu interessieren braucht.

Die Suchanfrage als graphisches Suchobjekt eignet sich hervorragend dazu, nach einer erfolgten Suche Zwischenergebnisse und -statistiken zu präsentieren.

Als weiteres Konzept soll das Arbeiten mit Korpora und *Teilkorpora* in COSMAS-II weiterhin unterstützt und erleichtert werden. Die Bildung von Teilkorpora oder *Arbeitskorpora* soll nun auch interaktiv über Suchanfragen geschehen können. Einerseits kann das Ergebnis einer Suche zur Defintion eines neuen Teilkorpus benutzt werden. Andererseits sollen auch bibliographische Kriterien die Dokumentauswahl bestimmen können. Da in COSMAS-II die bibliographischen Daten zu einem gespeicherten Dokument in seinem SGML-Kopfteil mitgegeben werden, wird die *bibliographische Suche* ein Sonderfall der allgemeinen Suche im neuen System sein und als solche mit der gleichen Anfragesprache formuliert werden können.

In Anlehnung an COSMAS-I ist die *dreistufige Ergebnisdarstellung* auch in die neue Version übernommen worden. Auf der obersten Ebene werden die Treffer nach Dokumenten zusammengefaßt und jedem Dokument die Trefferquote beigefügt. Auf der zweiten Ebene, auf welche man durch die Wahl eines Dokuments gelangt, werden die Treffer im Kontext einer einzigen Zeile als KWIC angezeigt. Auf der untersten Ebene wird der Treffer im Text dargestellt (größtmöglicher Kontext).

Eine SGML-Datenbank ist notwendigerweise mit dem Problem der *SGML-Datendarstellung* konfrontiert. Bei der reichhaltigen Palette an Möglichkeiten, die SGML-Strukturen bieten, muß eine erste Auswahl getroffen werden, da die verschiedenen Anwendungen und Benutzerbedürfnisse nicht alle antizipiert werden können, abgesehen davon, daß COSMAS-II in erster Linie für Recherchen eingesetzt wird und nicht, um SGML-Dokumente mit allen Spitzfindigkeiten zu verwalten und darzustellen. Wie schon erwähnt wurde, wird es gewisse Fälle geben, in denen SGML-Strukturen nicht angezeigt werden sollen, weil sie hauptsächlich dem Zweck dienen, komplexe Strukturen zu kodieren und in die Datenbank zu laden. Mit solchen Strukturen kommt der Benutzer bei der Abfrage nicht in Berührung, weil sie ihm, wie wir später sehen werden, entweder verborgen bleiben oder in einer benutzerfreundlicheren Form präsentiert werden. Es wird wiederum Fälle geben, in denen die Ergebnispräsentation mit einer Andeutung von SGML-Strukturen gewünscht wird. Damit ist gemeint, daß die SGML-Markierungen in einer simplen, nicht SGML-mäßig aktiven Form dargestellt werden sollen. Am anderen Ende der Darstellungskomplexität kann der Fall eintreten, in dem der Benutzer eine möglichst dokumentgetreue Darstellung der Treffer wünscht, das heißt, daß die SGML-Markierungen aktiv, also gemäß ihrer Bedeutung, angezeigt werden sollen, in welchem Falle die beste Lösung darin besteht, daß die Treffer direkt an einen angekoppelten SGML-Editor zur Darstellung übergeben werden. Schließlich soll auch die Möglichkeit bestehen, daß für die Ergebnisdarstellung eine eigene DTD entwickelt wird, mit deren Hilfe die Treffer in einer besonders anschaulichen Art und Weise markiert und exportiert werden können.

Ein letzter Schwerpunkt soll uns noch in diesem Kapitel interessieren, nämlich die *Nachverarbeitung von Ergebnismengen* unter Verwendung von *statistischen Analyseverfahren*. Verschiedene statistische Verfahren (siehe dazu u.a. Sinclair, 1995) in Form von unabhängigen Modulen werden nach dem Bausteinprinzip von COSMAS-II in das System integriert. Damit lassen sich verschiedene Aufgaben erfüllen, wie z.B. das Filtrieren und das

Reorganisieren der KWIC-Zeilen nach statischen Kriterien sowie das Gewinnen von neuen Informationen aus denselben. Diese Art der Nachverarbeitung erweist sich manchmal als notwendig, zumal sie bei großen Ergebnismengen erlaubt, relevante Ergebnisse zusammenzufassen und die Handarbeit zu beschleunigen.

3. Aspekte der Suchanfragesprache von COSMAS-II

In diesem Kapitel möchten wir vor allem einige Aspekte der Suchoperatoren und Suchobjekte von COSMAS-II vorstellen. Obwohl einige Suchoperatoren von der ersten Version übernommen worden sind, besteht nun in den *Suchobjekten* der zweiten Version ein wesentlicher Unterschied zur vorherigen. Unter Suchobjekten verstehen wir einerseits die *Textobjekte* oder *Textstellen*, die als Resultat einer jeden elementaren Suchanfrage oder als Kombination derselben erzeugt und zurückgeliefert werden, andererseits denjenigen Teil einer Suchanfrage, der sie spezifiziert. Die Einführung von SGML-Strukturen hat es nahe gebracht, daß Suchobjekte als zusammenhängende oder nichtzusammenhängende Textstellen zu betrachten sind. Es seien an dieser Stelle einige Beispiele angeführt, in welchen die verschiedenen Formen (oder Strukturen) von Suchobjekten, die das neue System zu verwalten imstande sein muß, gezeigt werden. Die Suchobjekte sind in den Beispielsätzen hervorgehoben worden:

Bsp. Aspekte der Abfragekomponente von COSMAS-II 1a) suche vollständige verbale Formen von "gehen"[1]:
... und **geht** ins Haus. Dies ist ...
Bsp. 1b) ... als sie auf ihr Zimmer **gegangen ist**. Warum ...
Bsp. 1c) ... und **ist** wie jeden Morgen zur Arbeit **gegangen**.

Das Suchobjekt in Bsp. 1c) ist hier also ein nichtzusammenhängendes.

Um diesem Umstand Rechnung zu tragen, sind die Suchoperatoren mit entsprechenden Erweiterungen neu definiert und die Suchobjekte der Suchanfragen mit verfeinerten Spezifikationsmöglichkeiten ausgestattet worden. Ein Suchobjekt ist nicht mehr rein nur eine Textstelle zwischen zwei Endpunkten, sondern eine Liste von Worttreffern und Wörtern. Diese können als *Eigenschaft* des Suchobjekts angesprochen werden, wie die nächsten Beispiele veranschaulichen:

Bsp. 2a) BEG: Erstes Wort eines Suchobjekts, **{ist}** in Bsp. 1c)
Bsp. 2b) END: Letztes Wort eines S., **{gegangen}** in Bsp. 1c)
Bsp. 2c) TRE: Alles Worttreffer eines S., **{ist, gegangen}** in Bsp. 1c)
Bsp. 2d) ALL: Alle Wörter zwischen BEG und END: **{ist, wie, jeden, Morgen, zur, Arbeit, gegangen}**in Bsp. 1c)
Bsp. 2e) [2]: Das 2. Wort in ALL, **{wie}** in Bsp. 1c)

[1]Es sei an dieser Stelle folgendes bemerkt: Ob eine Suche wie "verbale Form" zusammenhängende und nichtzusammenhängende Textstellen liefern kann, hängt davon ab, ob 1) die entsprechenden morphosyntaktischen Eigenschaften als Annotationen mit dem Korpus mitgeliefert wurden oder ob 2) ein entsprechendes Modul nachträglich in das COSMAS-II-Komplex eingebaut wird, das solche Eigenschaften detektiert und markiert.

Damit lassen sich nun die Abstandsoperatoren von COSMAS-I (Wort-, Satz- und Absatzabstand) auf Suchobjekte + Eigenschaften erweitern und durch neue Operatoren für die neuen Kombinationen ergänzen, wie z.B. die Operatoren IN (Suchobjekt1 soll IN Suchobjekt2 gesucht werden) oder ÜBERLAPP (Suchobjekt1 und Suchobjekt2 dürfen sich in einer zu spezifizierenden Weise überlappen).

Bsp. 3a) suche("Arbeit") /w2 suche(VerbaleForm("gehen", [BEG]))

In diesem Beispiel werden zwei Suchobjekte in einem Abstand von zwei Wörtern (Operator **/w2** in COSMAS-I) voneinander gesucht: einerseits das Wort "Arbeit", andererseits der erste Worttreffer der verbalen Form von "gehen".

Bsp. 3b) suche("Arbeit") IN suche(SGML-Element(Titel))

In diesem Beispiel wird das Wort "Arbeit" innerhalb eines Titels gesucht. Wie hier angedeutet, soll ein Titel die Textstelle sein, die durch eine entsprechende SGML-Markierung gekennzeichnet ist.

Die Abstandsoperatoren bergen in sich das Konzept der *minimalen* und *maximalen Gruppenbildung*, das sich auf die Art und Weise, wie Treffer eines Abstandsoperators zusammengefaßt werden, bezieht. Dies soll anhand des nächsten Beispiels veranschaulicht werden:

Bsp. 4) Suchanfrage: suche("um") /s0 suche(Infinitiv)
Korpus: "..., um auf diese Art das Feld zu räumen und zu verlassen."
Bsp. 4a) Treffer: "..., **um** auf diese Art das Feld zu **räumen** und zu **verlassen.**"
Bsp. 4b) Treffer1: "..., **um** auf diese Art das Feld zu **räumen** und zu verlassen."
Treffer2: "..., **um** auf diese Art das Feld zu räumen und zu **verlassen.**"

Je nachdem, ob man nur an den "maximalen" Treffer (Bsp. 4a), der alle die Suchanfrage erfüllenden Treffer zusammenfaßt, oder an die einzelnen Treffer (Bsp. 4b), bei denen ein Suchobjekt in der Suchanfrage je einer Textstelle zugeordnet wird, interessiert ist, soll man zwischen der maximalen und der minimalen Gruppenbildung wählen können.

Die Beispiele 1 bis 3 deuten schon an, daß auch Suchoperationen auf die SGML-Strukturen ausgeführt werden sollen. Die Suchoperationen auf dieselben beziehen sich auf die Elemente, auf deren Attribute und Attributswerte.

Bsp. 5a) Suche alle Textstellen, die durch das Element <s> (für Segment) markiert sind.
Bsp. 5b) Suche alle Textstellen, die durch das Element <s> mit Attribut *Typ* und Attributswert 'verbale Gruppe' gekennzeichnet sind. Also wird folgende Markierung gesucht: <s typ='verbale Gruppe'> ... *Text* ... </s>
Bsp. 5c) Suche alle Textstellen, die allein durch das Attribut *Typo* (für typographische Information) und den Wert *kursiv* gekennzeichnet sind, dabei soll es keine Rolle spielen, welche Elemente durchsucht werden.

Es gehört dazu, daß Suchoperationen auf SGML-Strukturen durch **eigene** (d.h. auf 'Elementebene' bezogene) Boolesche Operatoren miteinander verknüpft werden, wenn mehrere Bedingungen auf ein und derselben Struktur angegeben werden sollen.

Bsp. 5d) Suche alle Textstellen, die durch ein <div0> Element markiert sind, welches durch typo=fett **und** schrift=12 **und** einheit=punkt gekennzeichnet ist (<div0 typo=fett schrift=12 einheit=punkt>).

Suchoperationen auf SGML-Strukturen liefern aber genau so wie Abstandsoperatoren Suchobjekte zurück, die auf Textstellen verweisen. Somit kann man sie gleichermaßen in Abstandsoperatoren verwenden:

Bsp. 6) Suche alle Vorkommnisse von "gehen", die im gleichen Satz (Operator **/s0** in COSMAS-I) wie eine kursiv gedruckte Stelle erscheinen:

suche("gehen") /s0 suche(Attribut-Wert(typ,kursiv))

Die Abfragesprache ist zusätzlich mit einem *Namenreferenzierungsmechanismus* versehen, der es erlaubt, Teile einer Suchanfrage zu benennen und zu referenzieren:

Bsp. 7) (suche("gehen")**::G** /s0 suche("hinauf")**::H**)
UND (**Ref(G)** /w1 suche("häufig"))
UND (**Ref(H)** /w2 suche("hinunter"))

In diesem Beispiel ist dem Suchobjekt "gehen" der Name **G**, dem Suchobjekt "hinauf" der Name **H** gegeben worden, so daß sie später wieder verwendet werden können (durch **Ref(G)** und **Ref(H)**). Implizit ist in diesem Beispiel auch der selbstsprechende Boolesche Operator UND (analog ODER) verwendet worden.

Um dieses Kapitel abzuschließen, seien noch die Korpusbildungsoperatoren erwähnt. Aus einer *Treffer-* oder *Ergebnismenge* läßt sich ein Teilkorpus, bestehend aus der Liste aller Dokumente, die mindestens einen Treffer enthalten, definieren. Diese *Teilkorpusdefinition* kann benannt und zum *aktiven Arbeitskorpus* gemacht werden, auf welches alle nachfolgenden Arbeitsschritte angewandt werden sollen. Andererseits können auf eine Ergebnismenge auch eine oder mehrere Teilkorpusdefinitionen *appliziert* werden, um diese Menge nach korpusgebundenen Kriterien darzustellen.

4. Das Arbeiten mit graphischen Suchanfragen

Wenden wir uns nun der graphischen Komponente zu, so unterscheiden wir vier Hauptfenster. Die Abbildung 2 zeigt die Anfrageerstellung und die Korpusauswahl, die eng miteinander verwandt sind. Eine Anfrage kann in beiden Fenstern mit den gleichen graphischen Werkzeugen erstellt werden, wobei die Suche im vorderen Fenster auf die Texte, während sie im hinteren Fenster auf den bibliographischen Teil der Dokumente angewandt wird.

Beide Fenster enthalten die vier folgenden Bestandteile: *A* ist die Leiste, in welcher die zur Verfügung stehenden Suchoperatoren als Ikone präsentiert sind. *B* zeigt eine sich im Entstehen befindende graphische Suchanfrage, für welche der erste (obere) Operand durch einen weiteren Suchoperator ersetzt ist und der zweite (untere) noch frei ist. Die Leiste *C* enthält die schon erhaltenen Suchergebnisse als Ikone. Analog dazu werden in der Leiste *C* der Korpusauswahl die Ergebnisse automatisch in Korpusdefinitionenikonen umgewandelt. Eine interessante Interaktion zwischen beiden Fenstern besteht darin, daß Ergebnisse der

Textsuche ebenfalls in die Leiste der Korpusdefinitionen übertragen werden können, während Korpusdefinitionen auf die Ergebnismengen appliziert werden können, um diese auf spezielle Teilkorpora einzuschränken. Die Zeile *D* schließlich erlaubt die Visualisierung der aufgelösten graphischen Suchanfrage in Klammernotation und ist für erfahrene Benutzer direkt modifizierbar. Die graphische Suchanfrage *B* veranschaulicht außerdem nochmals die Konzeption einer Suche durch Einsetzen und Kombinieren von leeren Schablonen (oder Kästchen).

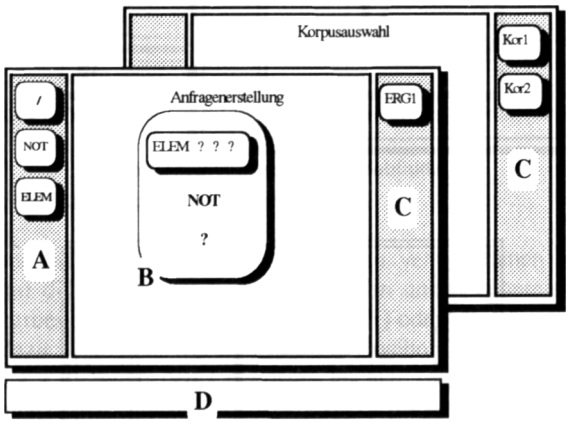

Abbildung 2: Suchanfragenerstellung und Korpusauswahl

Die Abbildung 3 zeigt einige der möglichen Ergebnisdarstellungsformen auf. Wieder ausgehend von der Anfrageerstellung (Fenster *E*) können die Zwischenergebnisse direkt in der graphischen Suchanfrage dargestellt werden. In der Abbildung 3 wird z.B. für jeden Operanden (d.h. für jede Teilsuche) die Anzahl der erhaltenen Treffer und Dokumente angezeigt. Im mittleren Fenster beginnt die dreistufige Ergebnispräsentation. Im Teil *F* sind die Dokumente mit ihren bibliographischen Angaben aufgelistet, im Teil *G* die Treffer als KWIC. Im dritten Fenster (*H*) schließlich wird jeder Treffer im Text angezeigt. Wenn nicht anders vom Benutzer erwünscht, beziehen sich die Ergebnisdarstellungen in den Teilen *F*, *G* und *H* auf das ganze Suchobjekt, aber sie könnten auch auf jeden Bestandteil der Suchanfrage in *E* eingeschränkt werden.

Wie im Kapitel 2 schon erwähnt wurde, kann eine Suchanfrage für den wiederholten Gebrauch als neues, *offenes* benutzerdefiniertes Ikon in die linke Leiste abgelegt werden, in welchem später beim Editieren alle Operanden modifiziert werden können. In der Abbildung 4 z.B. handelt es sich um eine Suche der Art CAT=VRB entsprechend einer Kodierung mit SGML/TEI und den für COSMAS-II typischen Operatoren für Elemente und Attribute.

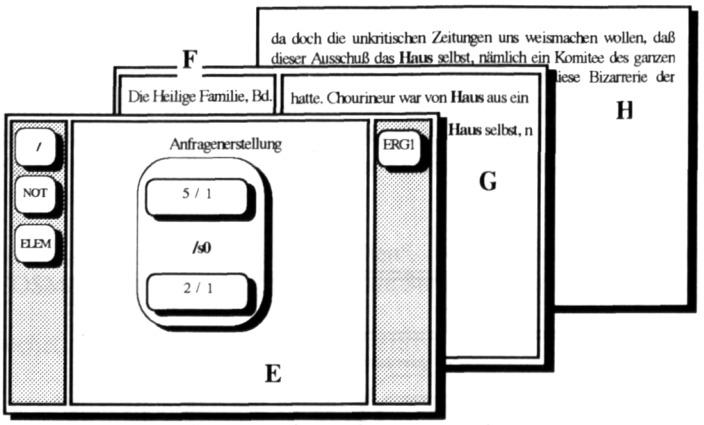

Abbildung 3: Ergebnispräsentation

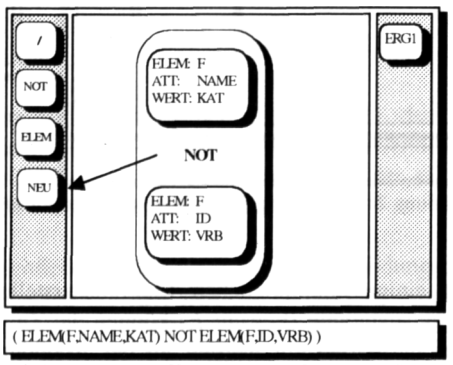

Abbildung 4: Inhalt einer offenen Suchanfrage und "Ikonisierung"

Will man die "Umständlichkeit" der SGML-Kodierung verbergen, so kann die Suchanfrage auch unter einer Maske zum Verschwinden gebracht werden, wobei die freiwählbaren Operanden (wie in Abbildung 5 dargestellt, in der nur noch ein freies Feld für die Eingabe der syntaktischen Kategorie besteht) in der Maske spezifiziert werden und diese als *geschlossenes* Ikon abgelegt wird. Als nützliche Konsequenz aus dieser Vorgehensweise ergibt sich automatisch, daß der Benutzer auch nicht mehr mit Daten und Zwischenergebnissen der verborgenen Strukturen konfrontiert wird. Da die Natur dieser Strukturen aber keineswegs verändert worden ist, können sie beim Wechsel in einen erfahrenen Modus unproblematisch wieder angezeigt werden.

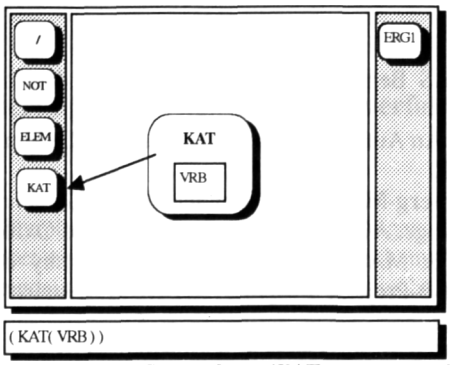

Abbildung 5: Verschlossene Suchanfrage (*KAT*) mit entsprechendem Ikon

5. Der Einsatz von statistischen Analyseverfahren

Der Einsatz von statistischen Analyseverfahren ist in verschiedenen Varianten denkbar. Prinzipiell zielt er auf die Gewinnung von neuen Daten und Erkenntnissen ausgehend von einem großen unverarbeiteten Textbestand (Kategorie 1) oder einer Ergebnismenge (Kat. 2), er kann aber auch dazu verwendet werden, größere Ergebnismengen so umzustrukturieren, daß ihre Auswertung z.B. durch den Linguisten leichter fällt (Kat. 3). Verfahren aus der ersten Kategorie können sich natürlich auf Korpusverwaltungssysteme wie COSMAS-I und COSMAS-II stützen, schöpfen aber die Datenzugriffskomponente eines Systems wie das eben beschriebene nur elementar aus.

Von den in COSMAS-II integrierten Verfahren fällt typischerweise das sogenannte *Collocate*-Programm in die zweite Kategorie. Ausgehend von einer KWIC-Repräsentation einer Ergebnismenge wird eine Liste von Wörtern (sogenannte *Kollokationen*) erstellt, welche sich mit einer gewissen Signifikanz in bezug auf das KWIC verhalten. Kollokationsberechnungen lassen sich auf Grund verschiedener Parameter wie +/- linker Kontext, +/- rechter Kontext, Größe des Kontextfensters, Häufigkeit der Kollokation und diverse zusätzliche Schwellwerte (auf Grund von statistischen Tests) bestimmen und erfüllen somit unterschiedliche Dienste. Z.B. wird am IDS bei der Untersuchung der Valenz von Nomen das kombinierte Verfahren aus Suche von bestimmten Nomen und Bestimmung der Kollokationen angewendet, um die gewünschten Valenzgruppen auf einfache Art zu extrahieren.

Aus der dritten Kategorie möchten wir das *Typical*-Programm erwähnen, das Konkordanzzeilen auf Grund der vorkommenden Kollokationen sortiert und sie auf Grund von ähnlichen Ausdrucksweisen automatisch zusammenfaßt.

Die in COSMAS-II integrierten Verfahren sind außerdem ein Versuch, zwei Konzepte zu verwirklichen, die ihre Anwendbarkeit für *multilinguale* Textdatenbanken wie diese erhöhen. Es handelt sich um die *Sprachunabhängigkeit* (Language Independency) und die *Annahme der minimalen Zusatzinformation* (Minimal Assumption). *Collocate* und *Typical* bedürfen keiner zusätzlichen Information und sind insofern sprachunabhängig, als daß sie nur Kenntnisse über die in der jeweiligen Sprache vorkommenden Alphabetkodierung voraussetzen.

Literatur:

al-Wadi Doris: COSMAS Benutzerhandbuch, Institut für deutsche Sprache, Mannheim, 1994

Bryan Martin: SGML - An Author's Guide to the Standard Generalized Markup Language. Addison-Wesley, 1988.

Burnard Lou & Sperberg-McQueen Michael (Ed.): Guidelines for Electronic Text Encoding and Interchange. ACH/ACL/ALLC, Chicago and Oxford, 1994.

Sinclair John M.:LIS & MAS, Progress Report. University of Birmingham, UK. In: MLAP93-21 MECOLB-Projekt, Progress Report II, August 94-January 95, Robert Neumann, Coordinator.